



09/288.256

国 特 許 庁
PATENT OFFICE
JAPANESE GOVERNMENT

別紙添付の書類に記載されている事項は下記の出願書類に記載されて
る事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed
th this Office.

出 願 年 月 日
Date of Application:

1998年 4月10日

出 願 番 号
Application Number:

平成10年特許願第114414号

出 願 人
Applicant (s):

株式会社リコー

RECEIVED

JUL 16 1999

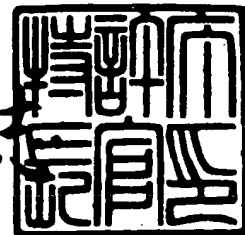
Group 2700

CERTIFIED COPY OF
PRIORITY DOCUMENT

1999年 4月23日

特 許 庁 長 官
Commissioner,
Patent Office

山 佐 建 彦



【書類名】 特許願

【整理番号】 9801993

【提出日】 平成10年 4月10日

【あて先】 特許庁長官殿

【国際特許分類】 G06F 17/21

【発明の名称】 文書分類装置、文書分類方法およびその方法をコンピュータに実行させるプログラムを記録したコンピュータ読み取り可能な記録媒体

【請求項の数】 7

【発明者】

 【住所又は居所】 東京都大田区中馬込 1丁目3番6号 株式会社リコー内

 【氏名】 長束 哲郎

【発明者】

 【住所又は居所】 東京都大田区中馬込 1丁目3番6号 株式会社リコー内

 【氏名】 宮地 達生

【発明者】

 【住所又は居所】 東京都大田区中馬込 1丁目3番6号 株式会社リコー内

 【氏名】 嶋田 敦夫

【発明者】

 【住所又は居所】 東京都大田区中馬込 1丁目3番6号 株式会社リコー内

 【氏名】 武谷 一寿

【発明者】

 【住所又は居所】 東京都大田区中馬込 1丁目3番6号 株式会社リコー内

 【氏名】 剣持 栄治

【発明者】

 【住所又は居所】 東京都大田区中馬込 1丁目3番6号 株式会社リコー内

 【氏名】 中島 明子

【発明者】

 【住所又は居所】 東京都大田区中馬込 1丁目3番6号 株式会社リコー内

【氏名】 山崎 真湖人

【発明者】

【住所又は居所】 東京都大田区中馬込 1 丁目 3 番 6 号 株式会社リコー内

【氏名】 藤田 克彦

【特許出願人】

【識別番号】 000006747

【氏名又は名称】 株式会社リコー

【代表者】 桜井 正光

【提出物件の目録】

【物件名】 明細書 1

【物件名】 図面 1

【物件名】 要約書 1

【書類名】 明細書

【発明の名称】 文書分類装置、文書分類方法およびその方法をコンピュータに実行させるプログラムを記録したコンピュータ読み取り可能な記録媒体

【特許請求の範囲】

【請求項 1】 文書の内容に基づいて文書の分類をおこなう文書分類装置において、

一つまたは複数の項目から構成された文書データを入力する入力手段と、

前記入力手段により入力された文書データを構成する前記項目を指定する指定手段と、

前記指定手段により指定された項目に対応するデータの内容となるように前記文書データを変換する変換手段と、

前記変換手段により変換された変換データをもちいて文書を分類する分類手段と、

を備えたことを特徴とする文書分類装置。

【請求項 2】 文書の内容に基づいて文書の分類をおこなう文書分類装置において、

一つまたは複数の項目から構成された文書データを入力する入力手段と、

前記入力手段により入力された文書データを構成する前記項目を指定する指定手段と、

前記指定手段により指定された項目に対応するデータの内容となるように前記文書データを変換する変換手段と、

前記変換手段により変換された変換データをもちいて各文書の特徴ベクトルを生成する文書ベクトル生成手段と、

前記文書ベクトル生成手段により生成された各文書の特徴ベクトルをもちいて文書を分類する分類手段と、

を備えたことを特徴とする文書分類装置。

【請求項 3】 前記変換手段は、前記文書データを変換する際、前記各項目のデータが分離可能となるように前記項目のデータ間に所定の記号を挿入することを特徴とする請求項 1 または 2 に記載の文書分類装置。

【請求項 4】 文書の内容に基づいて文書の分類をおこなう文書分類方法において、

一つまたは複数の項目から構成された文書データを入力する入力工程と、
前記入力工程により入力された文書データを構成する前記項目を指定する指定工程と、

前記指定工程により指定された項目に対応するデータのみの内容となるように前記文書データを変換する変換工程と、

前記変換工程により変換された変換データをもちいて文書を分類する分類工程と、

を含んだことを特徴とする文書分類方法。

【請求項 5】 文書の内容に基づいて文書の分類をおこなう文書分類方法において、

一つまたは複数の項目から構成された文書データを入力する入力工程と、
前記入力工程により入力された文書データを構成する前記項目を指定する指定工程と、

前記指定工程により指定された項目に対応するデータのみの内容となるように前記文書データを変換する変換工程と、

前記変換工程により変換された変換データをもちいて各文書の特徴ベクトルを生成する文書ベクトル生成工程と、

前記文書ベクトル生成工程により生成された各文書の特徴ベクトルをもちいて文書を分類する分類工程と、

を含んだことを特徴とする文書分類方法。

【請求項 6】 前記変換工程は、前記文書データを変換する際、前記各項目のデータが分離可能となるように前記項目のデータ間に所定の記号を挿入することを特徴とする請求項 4 または 5 に記載の文書分類方法。

【請求項 7】 前記請求項 4～6 のいずれか一つに記載された方法をコンピュータに実行させるプログラムを記録したことを特徴とするコンピュータ読み取り可能な記録媒体。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】

この発明は、文書の内容に基づいて文書を分類する文書分類装置、文書分類方法およびその方法をコンピュータに実行させるプログラムを記録したコンピュータ読み取り可能な記録媒体に関する。

【0002】

【従来の技術】

従来、文書分類装置として、たとえば、特開平7-36897号公報記載の文書分類装置には、文書を単語を特徴とする文書ベクトルとみなし、クラスタリング手法を用いてこれらの文書ベクトルを群分けし、文書の自動分類をおこなうものが記載されている。

【0003】

また、通常、文書データは一般的にデータベース化されており、文書内容だけでなく作成日や作成者などの書誌的項目が付加されていたり、また文書内容自体が複数の項目を含んでいる場合が多い。たとえば、特許公報は、「特許請求の範囲」「発明の詳細な説明」といった複数の項目から構成されている。

【0004】

【発明が解決しようとする課題】

しかしながら、上記従来技術の文書分類装置は、複数の項目を持つ文書データに対して、操作者が分類対象とする項目を任意に指定することができないことから、分類に悪影響を与えるデータが付加されていたり、また、複数の項目を組み合わせることが出来ないことから、分類に有効なデータが不足したりして、精度の高い分類結果を得ることができないという問題があった。

【0005】

この発明は、上述した従来例による問題点を解消するため、操作者の意図が反映された精度の高い分類をおこなうことができる文書分類装置、文書分類方法およびその方法をコンピュータに実行させるプログラムを記録したコンピュータ読み取り可能な記録媒体を提供することを目的とする。

【0006】

【課題を解決するための手段】

上述した課題を解決し、目的を達成するため、請求項1の発明に係る文書分類装置は、文書の内容に基づいて文書の分類をおこなう文書分類装置において、一つまたは複数の項目から構成された文書データを入力する入力手段と、前記入力手段により入力された文書データを構成する前記項目を指定する指定手段と、前記指定手段により指定された項目に対応するデータのみの内容となるように前記文書データを変換する変換手段と、前記変換手段により変換された変換データをもちいて文書を分類する分類手段と、を備えたことを特徴とする。

【0007】

この請求項1の発明によれば、文書を分類する際に、指定された項目の内容データだけが用いられるので、その他の項目の内容による分類結果への影響を防ぐことができる。そのため、操作者が期待する分類の観点に必要なと思われる文書データの項目を指定することにより、操作者が望む分類により近い精度の高い分類を効率よくおこなうことが可能である。

【0008】

また、請求項2に係る文書分類装置は、文書の内容に基づいて文書の分類をおこなう文書分類装置において、一つまたは複数の項目から構成された文書データを入力する入力手段と、前記入力手段により入力された文書データを構成する前記項目を指定する指定手段と、前記指定手段により指定された項目に対応するデータのみの内容となるように前記文書データを変換する変換手段と、前記変換手段により変換された変換データをもちいて各文書の特徴ベクトルを生成する文書ベクトル生成手段と、前記文書ベクトル生成手段により生成された各文書の特徴ベクトルをもちいて文書を分類する分類手段と、を備えたことを特徴とする。

【0009】

この請求項2の発明によれば、文書を分類するための各文書の特徴ベクトルを生成する際に、指定された項目の内容データだけが用いられるので、その他の項目の内容による分類結果への影響を防ぐことができる。そのため、操作者が期待する分類の観点に必要なと思われる文書データの項目を指定することにより、操作者が望む分類により近い精度の高い分類をおこなうことが可能である。

【0010】

また、請求項3に係る文書分類装置は、請求項1または2の発明において、前記変換手段が、前記文書データを変換する際、前記各項目のデータが分離可能となるように前記項目のデータ間に所定の記号を挿入することを特徴とする。

【0011】

この請求項3の発明によれば、各変換データの間区切りとなる記号を挿入するので、形態素解析等の解析処理をおこなう際に、各項目に対応するデータをそのまま結合させることにより変換データ全体として全く別の意味が構成されることを回避することが可能である。

【0012】

また、請求項4に係る文書分類方法は、文書の内容に基づいて文書の分類をおこなう文書分類方法において、一つまたは複数の項目から構成された文書データを入力する入力工程と、前記入力工程により入力された文書データを構成する前記項目を指定する指定工程と、前記指定工程により指定された項目に対応するデータの内容となるように前記文書データを変換する変換工程と、前記変換工程により変換された変換データをもちいて文書を分類する分類工程とを含んだことを特徴とする。

【0013】

この請求項4の発明によれば、文書を分類する際に、指定された項目の内容データだけが用いられるので、その他の項目の内容による分類結果への影響を防ぐことができる。そのため、操作者が自分が期待する分類の観点に必要なと思われる文書データの項目を指定することにより、操作者が望む分類により近い精度の高い分類をおこなうことが可能である。

【0014】

また、請求項5に係る文書分類方法は、文書の内容に基づいて文書の分類をおこなう文書分類方法において、一つまたは複数の項目から構成された文書データを入力する入力工程と、前記入力工程により入力された文書データを構成する前記項目を指定する指定工程と、前記指定工程により指定された項目に対応するデータの内容となるように前記文書データを変換する変換工程と、前記変換工

程により変換された変換データをもちいて各文書の特徴ベクトルを生成する文書ベクトル生成工程と、前記文書ベクトル生成工程により生成された各文書の特徴ベクトルをもちいて文書を分類する分類工程と、を含んだことを特徴とする。

【0015】

この請求項5の発明によれば、文書を分類するための各文書の特徴ベクトルを生成する際に、指定された項目の内容データだけが用いられるので、その他の項目の内容による分類結果への影響を防ぐことができる。そのため、操作者が自分が期待する分類の観点に必要なと思われる文書データの項目を指定することにより、操作者が望む分類により近い精度の高い分類をおこなうことが可能である。

【0016】

また、請求項6に係る文書分類方法は、請求項4または5の発明において、前記変換工程が、前記文書データを変換する際、前記各項目のデータが分離可能となるように前記項目のデータ間に所定の記号を挿入することを特徴とする。

【0017】

この請求項6の発明によれば、各変換データの間に区切りとなる記号を挿入するので、形態素解析等の解析処理の際に、複数の項目のデータを一つのデータとして混同して扱われることを回避できるとともに、各項目ごとの内容データが瞬時に識別することが可能である。

【0018】

また、請求項7の発明に係る記憶媒体は、請求項4～6に記載された方法をコンピュータに実行させるプログラムを記録したことで、そのプログラムを機械読み取り可能となり、これによって、請求項4～6の動作をコンピュータによって実現することが可能である。

【0019】

【発明の実施の形態】

以下に添付図面を参照して、この発明に係る文書分類装置、文書分類方法およびその方法をコンピュータに実行させるプログラムを記録したコンピュータ読み取り可能な記録媒体の好適な実施の形態を詳細に説明する。

【0020】

(実施の形態 1)

まず、この発明の実施の形態 1 による文書分類装置を構成する情報処理システム全体のハードウェア構成を説明する。図 1 は、実施の形態 1 による文書分類装置を構成する情報処理システム全体のハードウェア構成を示す説明図である。図 1 において、実施の形態 1 による文書分類装置を構成する情報処理システムは、サーバー/クライアント方式で構成されている。すなわち、サーバー 101 と複数のクライアント 102 がネットワーク 103 によって接続されている。

【0021】

クライアント 102 は、分類データの生成、サーバー 101 への指示、分類結果の表示などをおこなう。一方、クライアント 102 からの指示に従い、サーバー 101 は文書（テキスト）分類に関する処理を膨大な数値演算によりおこない、その処理の結果をクライアント 102 へ送る。より具体的には、サーバー 101 においては、テキスト分類処理がおこなわれ、クライアント 102 においては、分類データ生成、処理実行指示、テキスト分類結果表示等がおこなわれる。

【0022】

また、サーバー 101 とクライアント 102 との間のデータのやりとりはファイル共有という方法をもちいる。すなわち、分類処理にもちいるファイルをサーバー 101 上の共有フォルダに作成することにより両者はデータのやりとりをおこなう。したがって、クライアント 102 からはサーバー 101 の共有フォルダをネットワーク共有して利用することが可能である。

【0023】

つぎに、サーバー 101 およびクライアント 102 のハードウェア構成について説明する。図 2 は、実施の形態 1 による文書分類装置を構成する情報処理システムにおけるサーバー 101 をハードウェア的に示す説明図である。サーバー 101 は、たとえばワークステーション（WS）等がもちいられる。

【0024】

図 2 において、201 はサーバー 101 全体を制御する CPU を、202 はブートプログラム等を記憶した ROM を、203 は CPU 201 のワークエリアとして使用される RAM 203 を、204 は通信回線 205 を介してネットワーク

103に接続され、そのネットワーク103と内部のインターフェイスを司るインターフェイス（I/F）を、206はデータを記憶するディスク装置を示している。200は上記各部を結合させるためのバスを示している。

【0025】

そのほか、文書情報、画像情報、機能情報等を表示するディスプレイ208や、データを入力するためのキーボード209およびマウス210等が同様に接続されていてもよい。さらに、ディスク装置206には、クライアント102との間のデータのやりとりをするための共有フォルダ207が設けられている。

【0026】

また、図3は、実施の形態1による文書分類装置を構成する情報処理システムにおけるクライアント102をハードウェア的に示す説明図である。クライアント102は、たとえばパーソナルコンピュータ（PC）等がもちいられる。

【0027】

図3において、301はシステム全体を制御するCPUを、302はブートプログラム等を記憶したROMを、303はCPU301のワークエリアとして使用されるRAMを、304はCPU301の制御にしたがってHD（ハードディスク）305に対するデータのリード／ライトを制御するHDD（ハードディスクドライブ）を、305はHDD304の制御で書き込まれたデータを記憶するHDを、306はCPU301の制御にしたがってFD（フロッピーディスク）307に対するデータのリード／ライトを制御するFDD（フロッピーディスクドライブ）を、307はFDD306の制御で書き込まれたデータを記憶する着脱自在のFDを、308はドキュメント、画像、機能情報等を表示するディスプレイをそれぞれ示している。

【0028】

また、309は通信回線310を介してネットワーク103に接続され、そのネットワーク103と内部のインターフェイスを司るインターフェイス（I/F）を、311は文字、数値、各種指示等の入力のためのキーを備えたキーボードを、312はカーソルの移動や範囲選択、あるいは表示画面に表示されたアイコンやボタンの押下やウィンドウの移動やサイズの変更等をおこなうマウスを、3

13はOCR (Optical Character Reader) 機能を備えた画像を光学的に読み取るスキャナを、314は分類結果を含むデータの内容等を印刷するプリンタを、315は上記各部を結合するためのバスをそれぞれ示している。

【0029】

つぎに、実施の形態1による文書分類装置の機能的構成について説明する。図4は実施の形態1による文書分類装置の構成を機能的に示すブロック図である。図4において、文書分類装置は、入力部401と、指定部402と、変換部403と、変換データ記憶部404と、分類部405と、分類結果記憶部406を含む構成である。

【0030】

つぎに、各構成部についてその内容を詳細に説明する。なお、入力部401、指定部402、変換部403、変換データ記憶部404、分類部405、分類結果記憶部406は、ROM202または302、RAM203または303、あるいはディスク装置306またはハードディスク316等の記録媒体に記録されたプログラムに記載された命令に従ってCPU201または301等が命令処理を実行することにより、各部の機能を実現するものである。

【0031】

(入力部401)

入力部401は、文書データを入力するものであり、たとえば、キーボード209または311、OCR機能を備えたスキャナ313、またはネットワーク103を経由して文書や文書群を得ることができるI/F204または309等である。また、入力部401は、上記以外に文書データを取得することができるものであれば、それらすべてを含む。

【0032】

たとえば、文書データがデータベース化されている場合に、そのデータベースが記録された媒体を本実施の形態の文書分類装置に組み入れた場合も文書データの入力とする。さらに、入力した文書データを記憶する図示しない文書データ記憶部を含んでいてもよい。この文書データ記憶部は、たとえば大容量のメモリを

有するサーバー101のディスク装置206等であってもよい。

【0033】

ここで、文書とは、本実施の形態にあっては、自然言語で記述された一つ以上の文の集まりであり、それが分類対象となる場合はこれを文書という。具体的には、公開特許公報や特定の新聞記事も文書であり、また、請求項や特定の一文を取り出したものであっても、これを文書と見なすものである。

【0034】

(指定部402)

指定部402は、文書データの項目を指定するものである。指定部は、具体的には3つの処理から構成される。

【0035】

まず、入力部401により入力された文書データから項目を抽出する(第1処理)。項目を抽出する方法としては、あらかじめ所定の符号(たとえば、「[」]」等)が付されている項目を検索し、その項目を選択する等の方法がある。

【0036】

上記第1処理は、指定部402でおこなう代わりに、入力部401においておこなってもよい。すなわち、入力部401が文書データを入力する際に、あわせてその文書データの項目の抽出をおこなう。その抽出結果は文書データと対応付けられて上記文書データ記憶部に記憶される。この場合は、当該抽出結果をもちいることにより、指定部402においては上記第1処理は省略されることになる。また、データベースの種類によってはあらかじめ項目に関する情報を有しているものがあり、その項目に関する情報を利用することによっても、上記第1処理は省略される。

【0037】

つぎに、第1処理による項目の抽出結果、上記文書データ記憶部に記憶された上記抽出結果、または上記項目に関する情報等に基づいて、抽出された各項目がどのような項目であり、その項目に対応する内容はどのようなものであるかの一覧を操作者に提示する(第2処理)。提示の方法としては、ディスプレイ208

または308に項目のみを、あるいは項目とその項目に対応する内容の全部または一部を表示する方法等がある。

【0038】

項目のみを表示する方法としては、たとえば、項目名を文書データ中の出現順序に基づいて横書で縦一列になるように羅列して表示するといった方法がある。この場合、表示画面上の表示行数よりも項目数が多くなる場合は、折り返して縦二列以上で表示してもよく、また、縦一列で表示して、表示画面を縦方向にスクロールできるようにしてもよい。

【0039】

項目とその項目に対応する内容の全部または一部を表示する方法としては、たとえば、上述の項目のみを表する方法と同様に、項目名を文書データ中の出現順序に基づいて横書で縦一列になるように羅列して表示し、さらにその右側に項目名と対応して配置される位置に同じく横書でその内容を表示するといった方法がある。この場合、表示画面上の表示列数よりも内容のデータ量が多くなる場合は、表示画面を横方向にスクロールできるようにしてもよい。

【0040】

また、項目とその項目に対応する内容の全部または一部を表示する別の方法としては、項目名のみを表示し、項目名が表示されている領域にカーソルを移動させ、所定の操作（マウス210または312のボタンあるいはキーボード209または311等の所定キーの押下）により、内容のデータの全部または一部をポップアップして表示するようにしてもよい。

【0041】

つぎに、操作者の指示に従って、提示（表示）された項目の中から分類処理の対象となる項目を一つまたは二つ以上を同時に指定する（第3処理）。指定の方法としては、キーボード209または311やマウス210または312等のポインティングデバイスからの指定に関する指示信号に基づいて、提示されている項目の中から該当する項目を指定する。

【0042】

この際、項目の指定は一つであってもよく、また、二つ以上を同時に指定して

もよい。また、結合の形態を併せて指定することもできる。さらに、指定の順序により、データの変換後の内容データの配列順を指定するようにしてもよい。

【0043】

(変換部403)

変換部403は、入力された文書データを前記指定部402により指定された項目に対応するデータの内容となるように文書データを変換するものである。具体的には、文書データ中の指定された項目に対応するデータだけを抽出し、抽出されたデータのみからなる変換データへ変換するものである。

【0044】

変換データは、単にもとの文書データにおける指定された項目の順序で各項目に対応するデータを羅列することにより変換される場合のみならず、たとえば指定された項目のデータ内容を文字列として結合して指定された項目のデータ内容だけを含む変換データとすることや、項目の順序をもとの文書データ内における順序と異なる順序に入れ替えてからデータを結合するように変換してもよい。

【0045】

また、変換部403は、変換データにおける各項目のデータが分離可能となるように項目のデータ間に所定の分離記号601を挿入する。これにより、各項目に対応するデータの切れ目を瞬時に把握することができる。

【0046】

また、この分離記号601は、形態素解析等の自然言語解析をおこなう場合に特に重要である。各項目に対応するデータが文の体をなしている場合（文の終わりが句点で終わっている場合）は、この分離記号がなくても文と文の切れ目を判断することができるが、各項目に対応するデータが文の体をなしていない場合（箇条書きの文、文の途中で項目が変わる等の場合）は、そのままデータ同士を結合させると、項目によっては、全く別の意味が構成されてしまう場合がある。そのような場合を回避するためにこの分離記号601を挿入する。

【0047】

分離記号601は、一般的には、切れ目を表す「/（スラッシュ）」がもちいられるが、変換データ中に「/」が存在する場合には、データの「/」との混同

が生じるので、別の記号をもちいることができる。また、この記号を挿入するか否かについてキーボード 209 または 311 に所定のキーを割付け、そのキーを押下するごとに、あるいは表示画面上に所定のアイコンを表示させて、マウス 210 または 312 によりそのアイコンをクリックするごとに、分離記号 601 を挿入する／挿入しないを交互に設定するようにしてもよい。

【0048】

(変換データ記憶部 404)

変換データ記憶部 404 は、変換データを記憶する記憶部である。変換データ記憶部 404 としては、たとえば、サーバー 101 のディスク装置 206 またはクライアント側のハードディスク 305、またはフロッピーディスク 307 等、変換データの容量の違いあるいは用途の違いにより、それぞれ設定することが可能である。

【0049】

変換データ記憶部 404 には、項目の設定順序等を含む変換データのほか、前記分離記号 601 等も記憶される。変換データ記憶部 404 に記憶された変換データは、別の分類の際に用いる等、活用を図ることができる。

【0050】

(分類部 405)

分類部 405 は、変換部 403 により変換された変換データまたは変換データ記憶部 404 に記憶されている変換データの内容にしたがって自動的に分類する。分類部 405 については、たとえば特開平 7-36897 号公報に開示された「文書分類装置」など従来の文書分類方法を用いて文書を分類することができる。

【0051】

(分類結果記憶部 406)

分類結果記憶部 406 は、分類部 405 により分類された結果を記憶する記憶部である。分類結果記憶部 406 としては、変換データ記憶部 404 と同様に、たとえば、サーバー 101 のディスク装置 206 またはクライアント側のハードディスク 305、またはフロッピーディスク 307 等、変換データの容量の違い

あるいは用途の違いにより、それぞれ設定することが可能である。

【0052】

つぎに、文書データと文書データを変換した変換データの一例について説明する。図5は文書データとその変換データの一例を示す説明図である。図5において、文書群として特許公報群をもちいた場合であり、501は文書データの一例であり、502は変換データの一例である。

【0053】

文書データ501は、「出願番号」、「出願日」、「発明者」、「発明の名称」、「目的」、「構成」、「請求項1」、「従来技術」、「課題を解決するための手段」、「作用」、「実施例」、「発明の効果」等の項目が含まれている。

【0054】

従来の文書分類装置では各文書データをひとまとまりとして取り扱うので、複数の項目を含む文書データに対してはすべての項目の内容データが分類処理の対象となり、操作者が望む分類の観点に不必要、あるいは悪影響を与える項目も含まれる場合がある。

【0055】

本実施の形態においては、分類をおこなう操作者は自分が望む分類の観点に必要と思われる項目を1つ以上指定することができる。たとえば特許公報文書群の分類をおこなう際に、操作者が「発明の課題」に注目したい場合は、「目的」、「課題を解決するための手段」、「作用」、「発明の効果」を指定する。また、解決手段に注目したい場合は、「課題を解決するための手段」および「実施例」を指定することができる。分類の対象となる項目が指定されると、指定された項目に基づいて文書データを変換する。

【0056】

図5にあっては、操作者が「目的」、「課題を解決するための手段」、「作用」、「発明の効果」の項目を指定した場合において、指定された項目の内容データだけを含むように変換した場合の例である。

【0057】

変更データ502から明らかなように、「目的」の項目に対応するデータであ

る「履歴とともに対応する画面情報を記憶しておき・・・ことを目的とする。」と、「課題を解決するための手段」の項目に対応するデータである「上記目的を達成するために・・・表示する表示手段とを有する。」と、「作用」の項目に対応するデータである「以上の構成において、入力手続きより・・・表示するように動作する。」と、「発明の効果」の項目に対応するデータである「以上説明したよう日本発明によれば・・・再現できる効果がある。」とが結合して一つの文書を構成している。

【0058】

また、図6は、同一の文書データをもちいて、操作者が「目的」、「課題を解決するための手段」、「作用」、「発明の効果」を指定した場合において、指定された項目の内容データだけを含み、各項目のデータ間に分離記号601（「／」）を挿入するように変換した場合の例である。

【0059】

つぎに、実施の形態1による文書分類装置の一連の処理の手順について説明する。図7は実施の形態1による文書分類装置の一連の処理の手順を示すフローチャートである。

【0060】

図7のフローチャートにおいて、まず、入力部1は文書データの入力をおこなう（ステップS710）。また、指定部402は項目の指定をおこなう（ステップS720）。

【0061】

変換部403は、ステップS710において入力された文書データをステップS720において指定されて項目の内容になるように変換データへ変換する（ステップS730）。また、必要に応じて分離記号601を項目に対応するデータ間に挿入する（ステップS740）。変換された変換データは、分離記号データとともに変換データ記憶部404により記憶される（ステップS750）。

【0062】

上記ステップS730において変換された変換データあるいは上記ステップS750において変換データ記憶部404によって記憶された変換データに基づい

て、分類部 405 は文書の分類をおこなう（ステップ S780）。分類処理が終了後、分類処理の結果は分類結果記憶部 406 により記憶され（ステップ S790）、すべての処理は終了する。

【0063】

以上説明したように、実施の形態 1 によれば、指定された項目により文書データが変換データへ変換され、その変換データに基づいて文書の分類をおこなうので、その他の不要な項目の内容による分類結果への影響を抑制することができる。また、分離記号 601 の挿入により、変換データにおける結合された項目ごとのデータの識別ができ、かつ、項目間のデータの結合による内容の混同を回避することができる。

【0064】

（実施の形態 2）

さて、実施の形態 1 では、変換データをもちいて文書を分類したが、以下に説明する実施の形態 2 のように、変換データを用いて文書の特徴ベクトルを生成し、その特徴ベクトルを用いて文書を分類するようにしてもよい。

【0065】

まず、実施の形態 2 による文書分類装置の機能的構成について説明する。図 8 は、実施の形態 2 による文書分類装置の構成を機能的に示すブロック図である。図 8 において、実施の形態 1 の図 4 と同一のものに関しては同じ番号を付して、その説明を省略する。

【0066】

図 8 において、文書分類装置は、入力部 401 と、指定部 402 と、変換部 403 と、変換データ記憶部 404 と、分類部 405 と、分類結果記憶部 406 のほかに、文書ベクトル生成部 801 と、文書ベクトル記憶部 802 とを含む構成である。

【0067】

なお、文書ベクトル生成部 801 と文書ベクトル記憶部 802 は、他の構成部と同様に、ROM 202 または 302、RAM 203 または 303、あるいはディスク装置 306 またはハードディスク 316 等の記録媒体に記録されたプログ

ラムに記載された命令に従ってCPU 201または301等が命令処理を実行することにより、各部の機能を実現するものである。

【0068】

(文書ベクトル生成部 801)

文書ベクトル生成部 801 は、各文書の特徴ベクトルを生成する。文書の特徴ベクトルを生成するためには、文書データに対して形態素解析等の自然言語解析処理をおこなう必要がある。この自然言語解析処理は、図示しない文書解析部によって、各文書データについて各項目ごとおこなわれる。形態素解析は従来の形態素解析手法を用いることができる。

【0069】

文書ベクトル生成部 801 では各文書データに対して前記文書解析部によって得られた解析結果を用いて文書ベクトルを生成するものである。この際に指定部 402 によって指定された項目に関する解析結果のみに基づいて文書ベクトルの生成をおこなう。たとえば各文書データに対して指定部 402 で指定された項目の内容データから得られる特徴ベクトルだけを加算して文書ベクトルを生成することで、指定部 402 で指定された項目の内容データだけを反映した文書ベクトルを生成することができる。

【0070】

(文書ベクトル記憶部 802)

文書ベクトル記憶部 802 は、文書ベクトル生成部 801 によって生成された各文書の特徴ベクトルを記憶する記憶部である。文書ベクトル記憶部 802 においては同一文書であっても指定部 402 により指定される項目によっては、その文書の特徴ベクトルが異なってくるので、指定ごとにそれぞれ文書の特徴ベクトルを記憶する。分類部 405 による文書の分類をおこなう際には、あらかじめ文書ベクトル記憶部 802 によって記憶された上記文書の特徴ベクトルをもちいるので、効率よく文書の分類をおこなうことができる。

【0071】

文書ベクトル記憶部 802 としては、たとえば、サーバー 101 のディスク装置 206 またはクライアント側のハードディスク 305、またはフロッピーディ

スク 307等を、変換データの容量の違いあるいは用途の違いにより、それぞれ設定することが可能である。

【0072】

(分類部 405)

分類部 405は、変換部 403により変換された各文書の特徴ベクトル間の類似度に基づいて文書を分類するものである。具体的には、生成された分類対象データに対して、カイ自乗法の手法、判別分析の手法、およびクラスタ分析の手法等の分類手法を適用することで、文書分類をおこなうことができる。ここではベクトルデータが適用できる分類手法であれば、その手法は問わない。

【0073】

つぎに、実施の形態 2 による文書分類装置の一連の処理の手順について説明する。図 9 は実施の形態 2 による文書分類装置の一連の処理の手順を示すフローチャートである。図 9 のフローチャートにおいて、ステップ S 710～S 750までは、実施の形態 1 の図 7 のフローチャートと同一ステップなので、同一ステップ番号を付して、その説明は省略する。

【0074】

上記ステップ S 730において変換された変換データあるいは上記ステップ S 750において記憶された変換データに基づいて、文書ベクトル生成部 801は各文書の特徴ベクトルの生成をおこなう(ステップ S 760)。生成された各文書の特徴ベクトルは文書ベクトル記憶部 802により記憶される(ステップ S 770)。

【0075】

上記ステップ S 760において変換された変換データあるいはステップ S 770において記憶された変換データに基づいて、分類部 405は文書の分類がおこなう(ステップ S 780)。分類処理が終了後、分類処理の結果は分類結果記憶部 406により記憶され(ステップ S 790)、すべての処理は終了する。

【0076】

以上、実施の形態 2 によれば、指定された項目により文書データが変換データへ変換され、変換データに基づいて、各文書の特徴ベクトルの生成をおこなうの

で、操作者の意図をより反映した文書の特徴ベクトルを用いて文書の分類をおこなうことができ、その他の不要な項目の内容による分類結果への影響を抑制することができる。

【0077】

【発明の効果】

以上説明したように、請求項1の発明によれば、文書を分類する際に、指定された項目の内容データだけが用いられるので、その他の項目の内容による分類結果への影響を防ぐことができる。そのため、操作者が期待する分類の観点に必要なと思われる文書データの項目を指定することにより、操作者が望む分類により近い精度の高い分類を効率よくおこなうことが可能な文書分類装置が得られるという効果を奏する。

【0078】

また、請求項2の発明によれば、文書を分類するための各文書の特徴ベクトルを生成する際に、指定された項目の内容データだけが用いられるので、その他の項目の内容による分類結果への影響を防ぐことができる。そのため、操作者が期待する分類の観点に必要なと思われる文書データの項目を指定することにより、操作者が望む分類により近い精度の高い分類をおこなうことが可能な文書分類装置が得られるという効果を奏する。

【0079】

また、請求項3の発明によれば、各変換データの間に区切りとなる記号を挿入するので、形態素解析等の解析処理の際に、複数の項目のデータを一つのデータとして混同して扱われることを回避できるとともに、各項目ごとの内容データが瞬時に識別することが可能な文書分類装置が得られるという効果を奏する。

【0080】

また、請求項4の発明によれば、文書を分類する際に、指定された項目の内容データだけが用いられるので、その他の項目の内容による分類結果への影響を防ぐことができる。そのため、操作者が自分が期待する分類の観点に必要なと思われる文書データの項目を指定することにより、操作者が望む分類により近い精度の高い分類をおこなうことが可能な文書分類方法が得られるという効果を奏する。

【0081】

また、請求項5の発明によれば、文書を分類するための各文書の特徴ベクトルを生成する際に、指定された項目の内容データだけが用いられるので、その他の項目の内容による分類結果への影響を防ぐことができる。そのため、操作者が自分が期待する分類の観点に必要なと思われる文書データの項目を指定することにより、操作者が望む分類により近い精度の高い分類をおこなうことが可能な文書分類方法が得られるという効果を奏する。

【0082】

また、請求項6の発明によれば、各変換データの間に区切りとなる記号を挿入するので、形態素解析等の解析処理の際に、複数の項目のデータを一つのデータとして混同して扱われることを回避できるとともに、各項目ごとの内容データが瞬時に識別することが可能な文書分類方法が得られるという効果を奏する。

【0083】

また、請求項7の発明によれば、請求項4～6のいずれか一つに記載された方法をコンピュータに実行させるプログラムを記録したことで、そのプログラムを機械読み取り可能となり、これによって、請求項4～6の動作をコンピュータによって実現することが可能な記録媒体が得られるという効果を奏する。

【図面の簡単な説明】

【図1】

この発明の実施の形態1による文書分類装置を構成する情報処理システム全体のハードウェア構成を示す説明図である。

【図2】

実施の形態1による文書分類装置を構成する情報処理システムにおけるサーバーをハードウェア的に示す説明図である。

【図3】

実施の形態1による文書分類装置を構成する情報処理システムにおけるクライアントをハードウェア的に示す説明図である。

【図4】

実施の形態1による文書分類装置の構成を機能的に示すブロック図である。

【図 5】

実施の形態 1 による文書分類装置における文書データおよび変換データの内容の一例を示す説明図である。

【図 6】

実施の形態 1 による文書分類装置における文書データおよび変換データの内容の別の一例を示す説明図である。

【図 7】

実施の形態 1 による文書分類装置の一連の処理の手順を示すフローチャートである。

【図 8】

この発明の実施の形態 2 による文書分類装置の構成を機能的に示すブロック図である。

【図 9】

実施の形態 2 による文書分類装置の一連の処理の手順を示すフローチャートである。

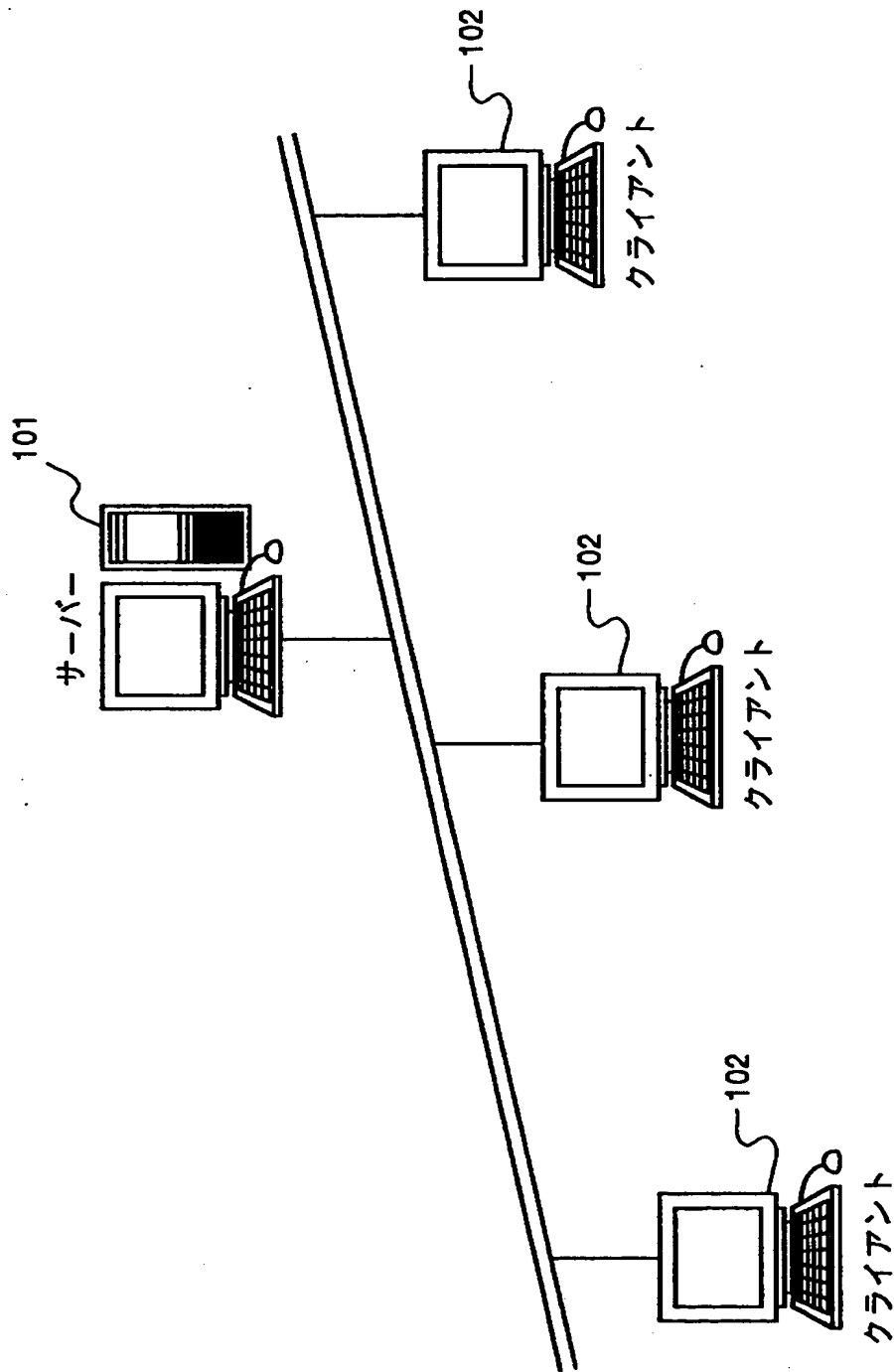
【符号の説明】

- 101 サーバー
- 102 クライアント
- 103 ネットワーク
- 201 CPU
- 204 I/F
- 206 ディスク装置
- 301 CPU
- 306 ハードディスク
- 308 ディスプレイ
- 309 I/F
- 311 キーボード
- 312 マウス
- 313 スキャナ

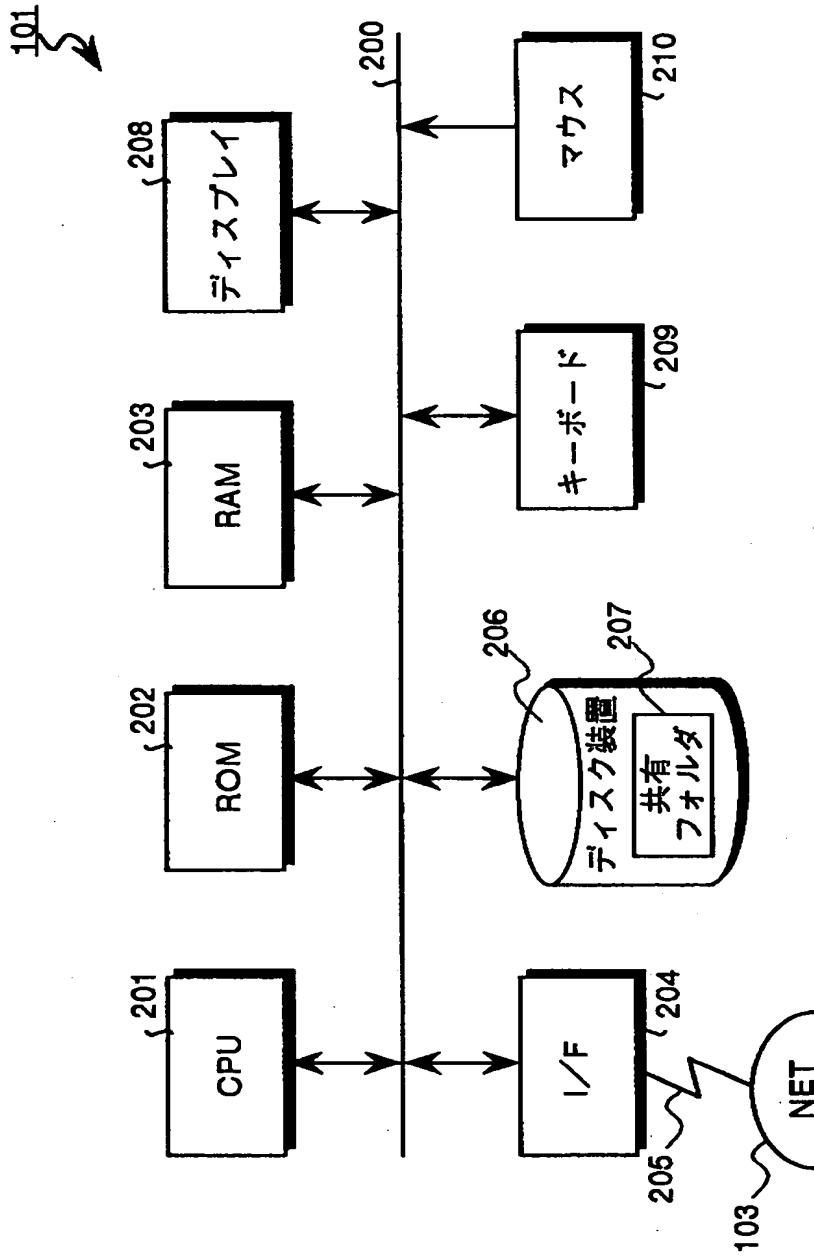
- 401 入力部
- 402 指定部
- 403 変換部
- 404 変換データ記憶部
- 405 分類部
- 406 分類結果記憶部
- 501 文書データ
- 502 変換データ
- 601 分離記号
- 801 文書ベクトル生成部
- 802 文書ベクトル記憶部

【書類名】 図面

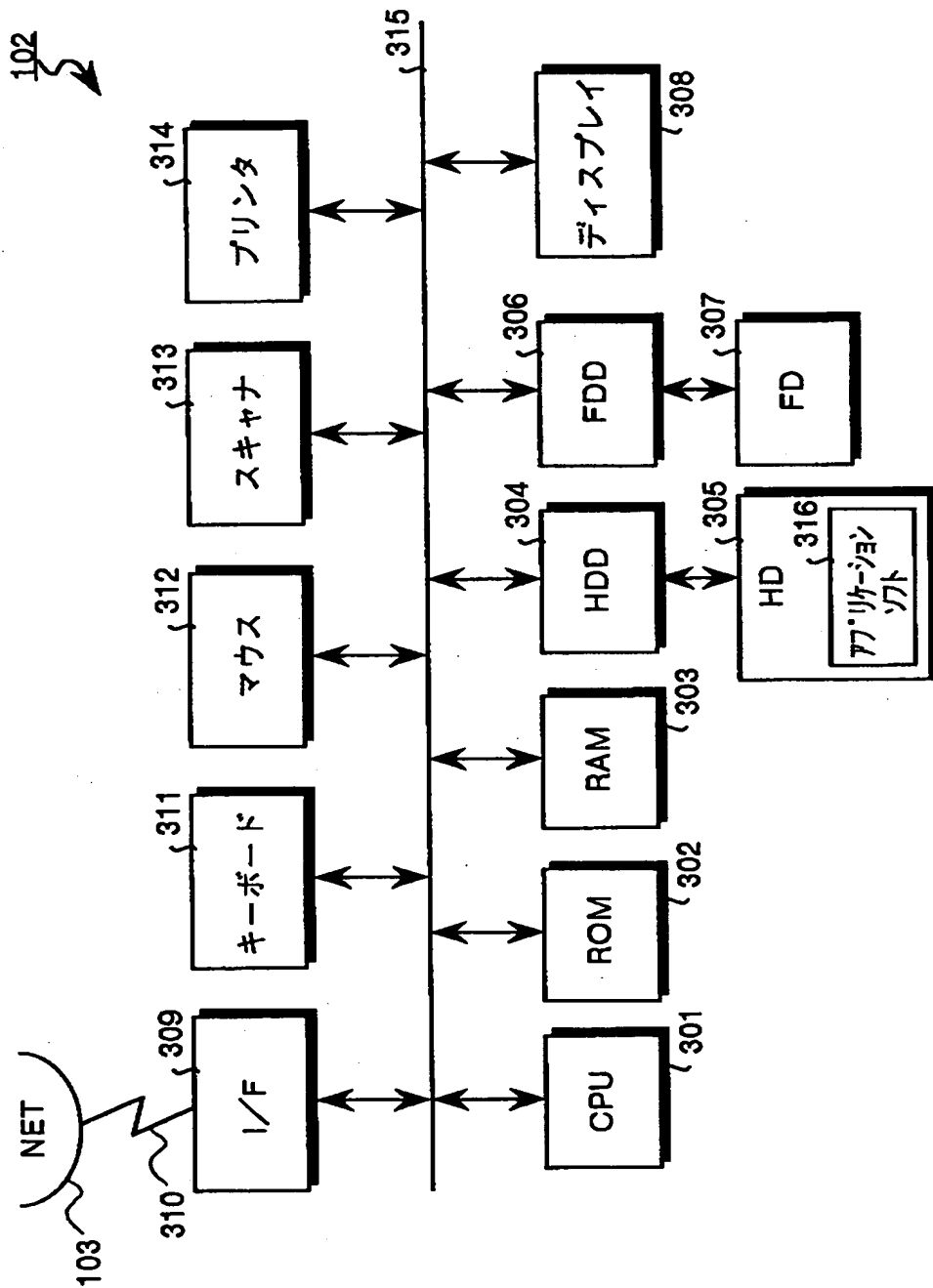
【図 1】



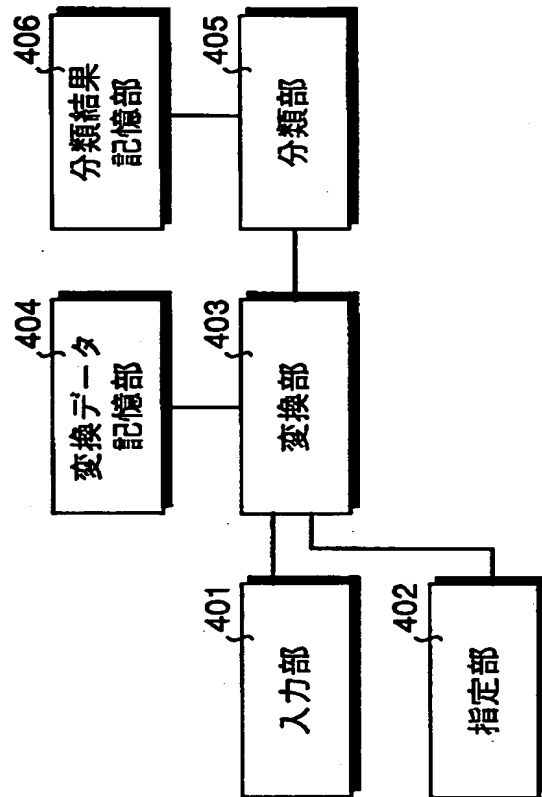
【図 2】



【図 3】



【図 4】



【図 5】

501

出願番号 特願平3-000000
出願日 平成3年(1991)〇月〇日
発明者 山田太郎
発明の名称 情報処理装置
目的 履歴とともに対応する画面情報を記憶しておき・・・ことを目的とする。
構成 入力部7より入力された・・・表示部24に表示される
請求項1 マルチウィンドウを表示して・・・特徴とする情報処理装置
従来技術 図2は、一般的な・・・表示することができる。
課題を解決するための手段 上記目的を達成するために・・・表示する表示手段とを有する。
作用 以上の構成において、入力手段より・・・表示するよう動作する。
実施例 以下、添付図面を参照して・・・ことが可能になる。
発明の効果 以上説明したように本発明によれば、・・・再現できる効果がある。 ...



502

履歴とともに対応する画面情報を記憶しておき・・・ことを目的とする。上記目的を達成するために・・・表示する表示手段とを有する。以上の構成において、入力手段より・・・表示するよう動作する。以上説明したように本発明によれば・・・再現できる効果がある。
--

【図 6】

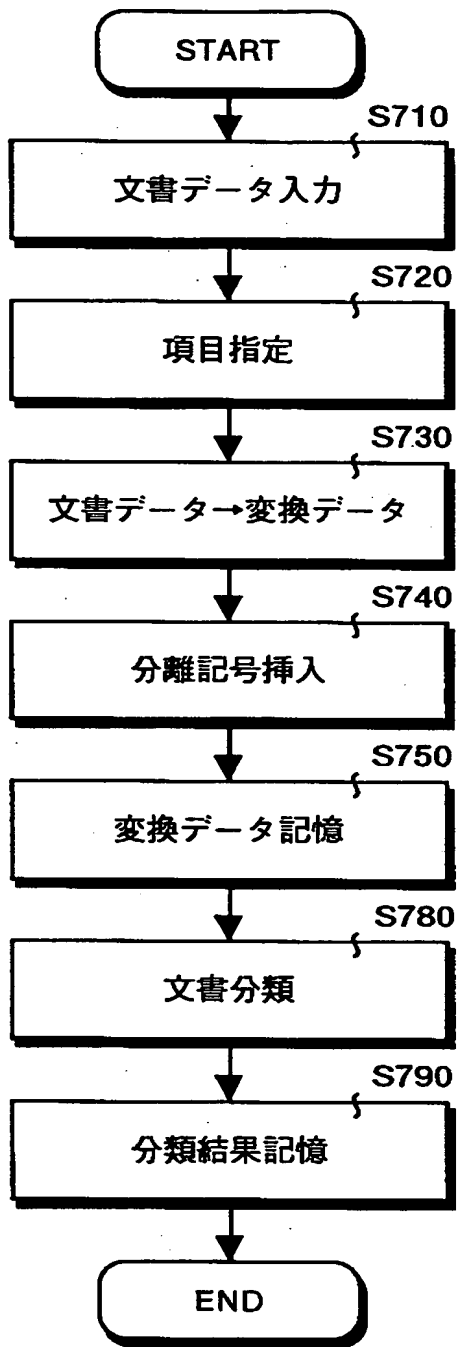
出願番号	特願平3-000000
出願日	平成3年(1991)〇月〇日
発明者	山田太郎
発明の名称	情報処理装置
目的	履歴とともに対応する画面情報を記憶しておき・・・ことを目的とする。
構成	入力部7より入力された・・・表示部24に表示される
請求項1	マルチウィンドウを表示して・・・特徴とする情報処理装置
従来技術	図2は、一般的な・・・表示することができる。
課題を解決するための手段	上記目的を達成するために・・・表示する表示手段とを有する。
作用	以上の構成において、入力手段より・・・表示するように動作する。
実施例	以下、添付図面を参照して・・・ことが可能になる。
発明の効果	以上説明したように本発明によれば、・・・再現できる効果がある。
...	



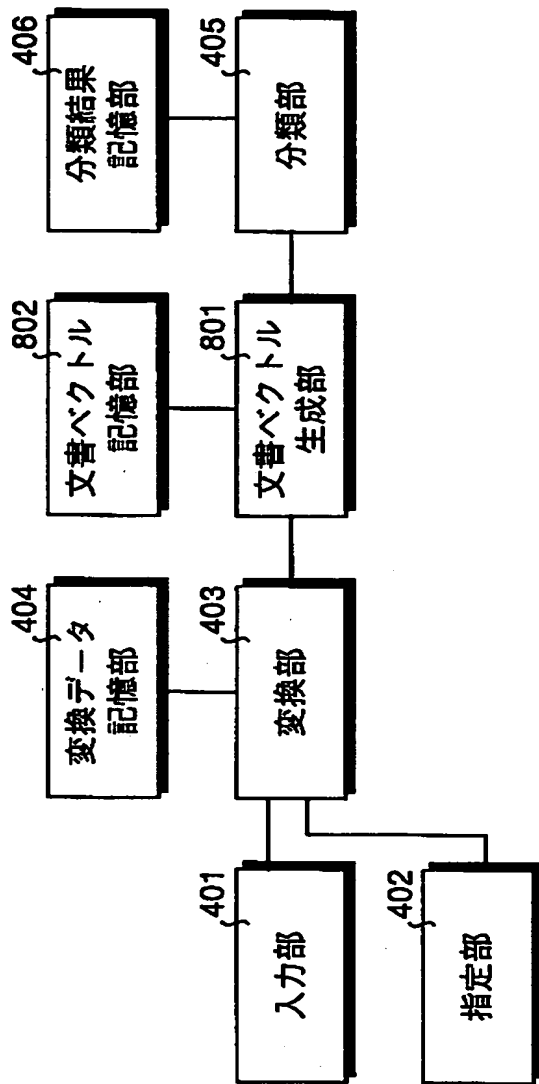
601

履歴とともに対応する画面情報を記憶しておき・・・ことを目的とする。○上記目的を達成するために・・・表示する表示手段とを有する。○以上の構成において、入力手段より・・・表示するように動作する。○以上説明したように本発明によれば・・・再現できる効果がある。

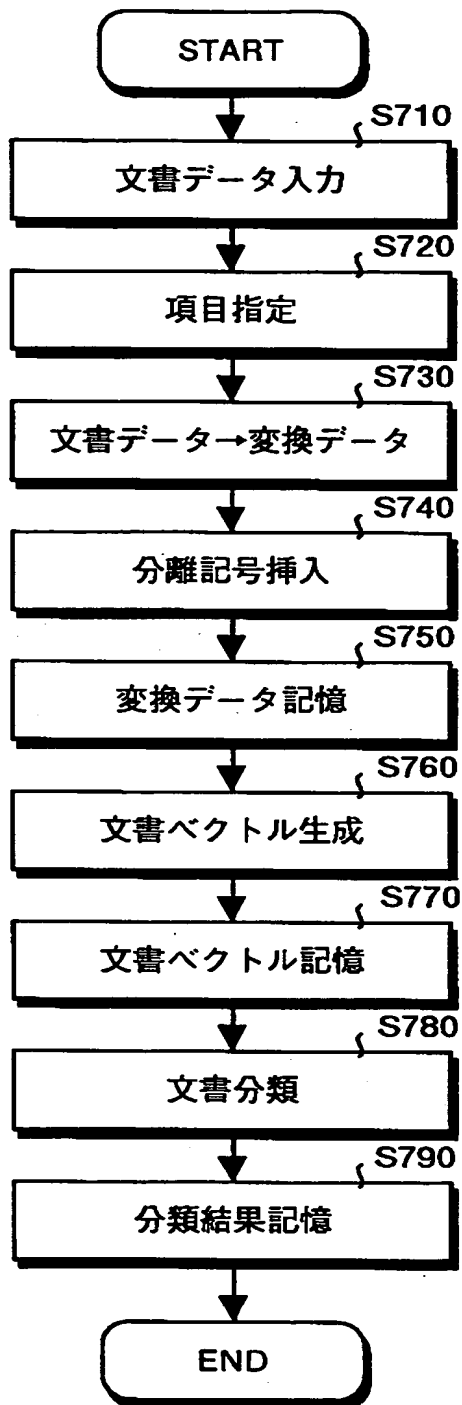
【図 7】



【図 8】



【図 9】



【書類名】 要約書

【要約】

【課題】 文書間の類似性に基づいて文書分類をおこなう際、操作者の意図を反映する文書分類をおこなうことを課題とする。

【解決手段】 一つまたは複数の項目から構成された文書データを入力する入力部401と、入力された文書データを構成する前記項目を指定する指定部402と、指定された項目に対応するデータのみの内容となるように前記文書データを変換する変換部403と、変換された変換データをもちいて文書を分類する分類部405とを備える。

【選択図】 図4

【書類名】
【訂正書類】

職権訂正データ
特許願

<認定情報・付加情報>

【特許出願人】

申請人

【識別番号】

000006747

【住所又は居所】

東京都大田区中馬込1丁目3番6号

【氏名又は名称】

株式会社リコー

出 願 人 履 歴 情 報

識別番号 [000006747]

1. 変更年月日	1990年 8月24日
[変更理由]	新規登録
住 所	東京都大田区中馬込1丁目3番6号
氏 名	株式会社リコー